

# Multimodel SuperEnsemble technique for quantitative precipitation forecasts in Piemonte region

D. Cane and M. Milelli

Regional Environmental Protection Agency – Arpa Piemonte, Torino, Italy

Received: 12 February 2009 – Revised: 3 November 2009 – Accepted: 26 January 2010 – Published: 12 February 2010

**Abstract.** The Multimodel SuperEnsemble technique is a powerful post-processing method for the estimation of weather forecast parameters reducing direct model output errors. It has been applied to real time NWP, TRMM-SSM/I based multi-analysis, Seasonal Climate Forecasts and Hurricane Forecasts. The novelty of this approach lies in the methodology, which differs from ensemble analysis techniques used elsewhere.

Several model outputs are put together with adequate weights to obtain a combined estimation of meteorological parameters. Weights are calculated by least-square minimization of the difference between the model and the observed field during a so-called training period. Although it can be applied successfully on the continuous parameters like temperature, humidity, wind speed and mean sea level pressure, the Multimodel SuperEnsemble gives good results also when applied on the precipitation, a parameter quite difficult to handle with standard post-processing methods. Here we present our methodology for the Multimodel precipitation forecasts, involving a new accurate statistical method for bias correction and a wide spectrum of results over Piemonte very dense non-GTS weather station network.

## 1 Introduction

Quantitative precipitation forecast (QPF) is one of the more difficult tasks in the weather forecasts, and any attempt in evaluating the uncertainty of the precipitation forecast contributes to a better characterization of the weather effects on the territory. The uncertainty in the QPF evaluation propagates down to the hydrological evaluation of the discharge

forecast in small and medium-sized catchments that are typical of the Mediterranean area.

From a given forecast dataset obtained from different models, many ensemble techniques can be applied, like the “poor man” ensemble (Eq. 1), a simple average of the models (requiring no training period), or a bias corrected ensemble (Eq. 2).

$$S = \frac{1}{N} \sum_{i=1}^N F_i \quad (1)$$

$$S = \bar{O} + \frac{1}{N} \sum_{i=1}^N (F_i - \bar{F}_i) \quad (2)$$

As suggested by the name, the Multimodel SuperEnsemble method (Krishnamurti et al., 1999, 2000, 2002; Williford et al., 2002) requires several model outputs, which are weighted with an adequate set of weights calculated during the so-called training period (Eq. 3).

$$S = \bar{O} + \sum_{i=1}^N a_i (F_i - \bar{F}_i) \quad (3)$$

where  $N$  is the number of models,  $a_i$  are the SuperEnsemble weights,  $F_i$  is the forecast value,  $\bar{F}_i$  is the mean forecast value in the training period and  $\bar{O}$  is the mean observation in the training period.

The calculation of the parameters  $a_i$  is given by the minimisation of the mean square deviation

$$G = \sum_{k=1}^T (S_k - O_k)^2 \quad (4)$$

where  $T$  is the training period length (days). By derivation



Correspondence to: D. Cane  
(daniele.cane@arpa.piemonte.it)

we obtain a set of  $N$  equations:

$$\begin{pmatrix} \sum_{k=1}^T (F_{1k} - \bar{F}_1)^2 & \sum_{k=1}^T (F_{1k} - \bar{F}_1)(F_{2k} - \bar{F}_2) & \dots & \sum_{k=1}^T (F_{1k} - \bar{F}_1)(F_{Nk} - \bar{F}_N) \\ \sum_{k=1}^T (F_{2k} - \bar{F}_2)(F_{1k} - \bar{F}_1) & \sum_{k=1}^T (F_{2k} - \bar{F}_2)^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{k=1}^T (F_{Nk} - \bar{F}_N)(F_{1k} - \bar{F}_1) & \dots & \dots & \sum_{k=1}^T (F_{Nk} - \bar{F}_N)^2 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^T (F_{1k} - \bar{F}_1)(O_k - \bar{O}) \\ \vdots \\ \vdots \\ \sum_{k=1}^T (F_{Nk} - \bar{F}_N)(O_k - \bar{O}) \end{pmatrix} \quad (5)$$

We then solve these equations using the Gauss-Jordan method.

In this paper, we will focus on the quantitative precipitation forecast, and in detail, we firstly examine the ensemble properties (Sect. 2), then we introduce a modification to the Multimodel technique (Sect. 3) that we consider appropriate for this variable. In Sect. 4 we describe the results obtained with or without this correction (Sect. 4.1), the comparison among Multimodel SuperEnsemble, Ensemble, and Poor Man Ensemble (Sect. 4.2) and the performance of the operational version of the technique which is running daily at ARPA Piemonte (Sect. 4.3). Eventually in Sect. 5 we draw some conclusion.

## 2 Ensemble evaluation

The ensemble approach and the related weighting techniques can be easily applied on variables like temperature or wind speed (Cane and Milelli, 2005, 2006), but a certain attention must be put in applying Multimodel SuperEnsemble on a variable like precipitation. There are in fact theoretical and empirical evidences to support the fact that precipitation, as the atmospheric turbulence, has a so-called multi-fractal behaviour (see for instance Lovejoy et al., 1996). This implies that similar features are observed in precipitation fields on a continuum of spatial scales from the very small (centimetres) to the very large (kilometres). This also implies that the numerical models bring along the quantitative precipitation forecast not only their intrinsic error, but also the uncertainty due to the stochastic properties of the precipitation field. Therefore, it is important to study firstly the characteristics of the data ensemble itself, without the application of any weight.

A standard technique for the evaluation of the reliability of an ensemble (that is whether the observed probability distribution is well represented by the ensemble) is the rank histogram (see for example Anderson, 1996 or Hamill, 2001), obtained by counting the rank of observations with respect to the ensemble member values. Precipitation forecast has to be treated with care, because the value is often equal to zero (i.e. no precipitation is forecasted or observed). In these

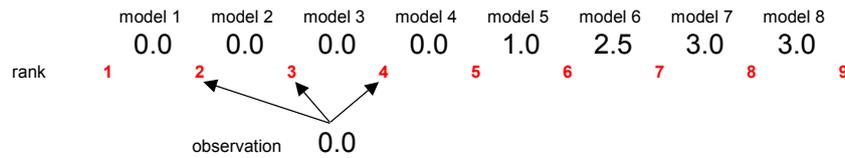
cases, we can find a certain number of equal forecasts and the ranking of the observation can be among one of these equal values. Considering the example in Fig. 1: the rank of the observation can be assigned, is not unique. If we choose one of these ranks systematically, we could have an incorrect calculation of the rank histogram. We then applied a random choice to all the cases where the rank is not uniquely determined, in order not to force the data towards an artificial distribution.

Arpa Piemonte manages a wide non-GTS weather station network. For this work, we used the data collected in the period March 2006–August 2008 from 342 stations. The data were averaged over the 13 warning areas designed by ARPA Piemonte in collaboration with the Civil Protection Department (Fig. 2) on a 6-h basis up to +72 h and for each of them the maximum values of observed precipitation has been assigned. Each warning areas so identified contains on average 26 stations, with a minimum of 11 and a maximum of 39. The chosen spatial scale is practical for the use in an alert system in medium- and large-scale catchments, but it is too much coarse for the discharge calculations in the smallest catchments. Finer resolution calculation are still under investigation.

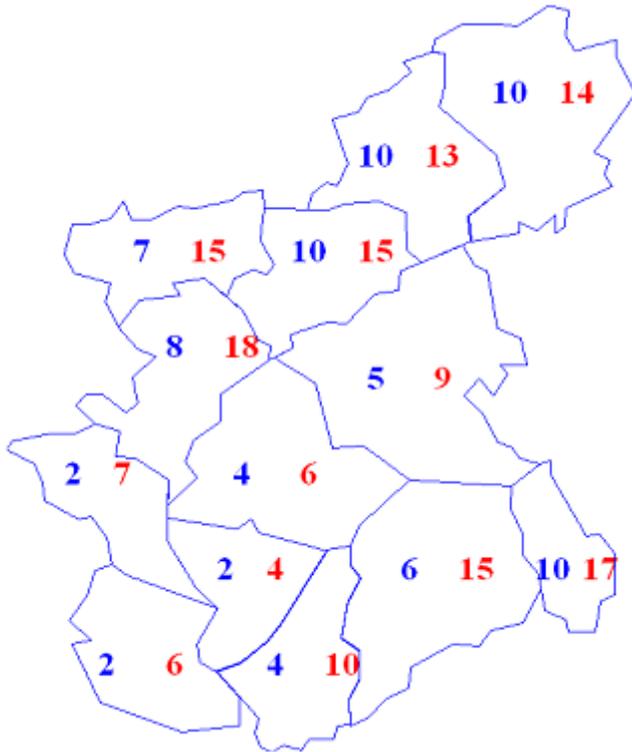
The models used in this research work are the ECMWF IFS global model (resolution:  $0.25^\circ$ ) and the  $0.0625^\circ$  resolution limited area models of the COSMO Consortium covering North-West Italy: COSMO-I7 (developed by USAM, ARPA-SIM, ARPA Piemonte), COSMO-EU (Deutscher Wetterdienst) and COSMO-7 (MeteoSwiss). It has to be highlighted that our operational implementation (see Sect. 4.3) is based on the forecasts given by the ECMWF model and by the Italian COSMO model only (see [www.cosmo-model.org](http://www.cosmo-model.org) for a more comprehensive overview of the Consortium activities and developments). The model forecasts are assigned to the same warning areas by taking the average and maximum values of the gridpoints falling into the given area. (ECMWF model:  $\sim 5$  points/warning area; COSMO models:  $\sim 56$  points/warning area)

In Fig. 3 we show the rank histograms of 8 runs of the 4 models in the period September 2006–August 2008 over all the Piemonte warning areas as described in Fig. 3. The ensemble is in general over-dispersed, that is the ensemble spread is too large, but it has to be pointed out that the inclusion of correct negatives (i.e. days of correct zero precipitation forecast) may artificially change the properties of the ensemble.

If we remove all the correct negatives, although keeping the false alarms and misses, we obtain a different rank histogram (Fig. 4), telling us more information: the ensemble usually overestimates the observation (the rank histogram peaks at low ranks), but there is a significant amount of underestimated forecasts (comprising misses), as evidenced by the higher value of the last rank column. These features are more accentuated for the maxima forecasts over the warning areas.



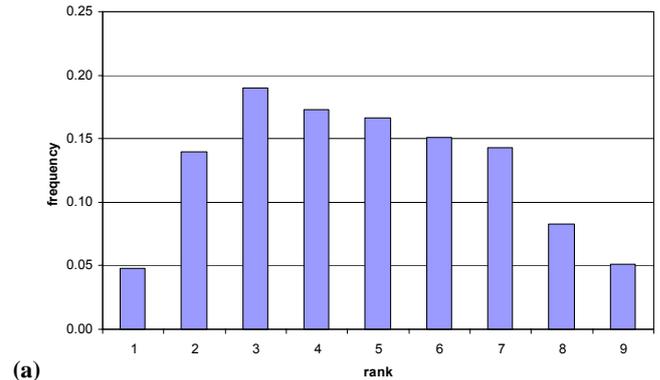
**Fig. 1.** Example of calculation of the rank of observation with respect to an ensemble of forecasts: the observation can not be assigned univocally.



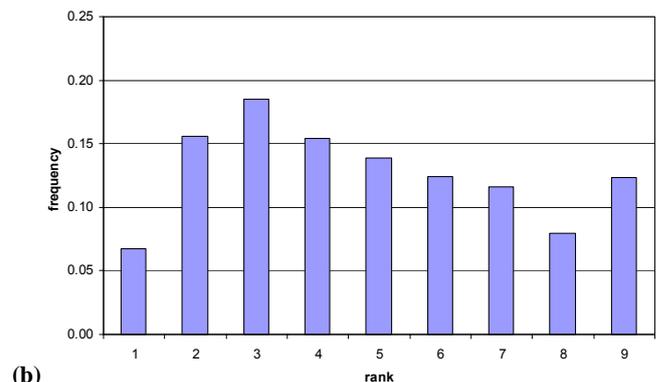
**Fig. 2.** A drawing of the 13 warning areas in Piemonte region. Blue numbers represent the observed (or forecast) 6-h precipitation (mm) averaged over the catchment, while red numbers are the maximum precipitation observed (or forecast) in the catchment.

We evaluated also the scatter plots of the ensemble mean Root Mean Square Error vs the ensemble spread (spread-skill diagram) in Fig. 5: the linear relationship between these two values is quite clear, and in particular the correlation coefficient for average values is quite high.

The rank histograms tell us that 88% of observations for average values and 80% for maxima already fall into the uncalibrated ensemble, while the scatter plot allows us to search for a relationship between the ensemble means to optimize the ensemble mean value. In both cases, the plots have been obtained adding up in forecast time (up to +72 h) and in space (over the 13 warning areas, see Fig. 5). The message that is evident from this preliminary analysis is that the ensemble has skill, but is not calibrated, therefore some statistical post-processing is necessary. The chosen method, as explained



(a)



(b)

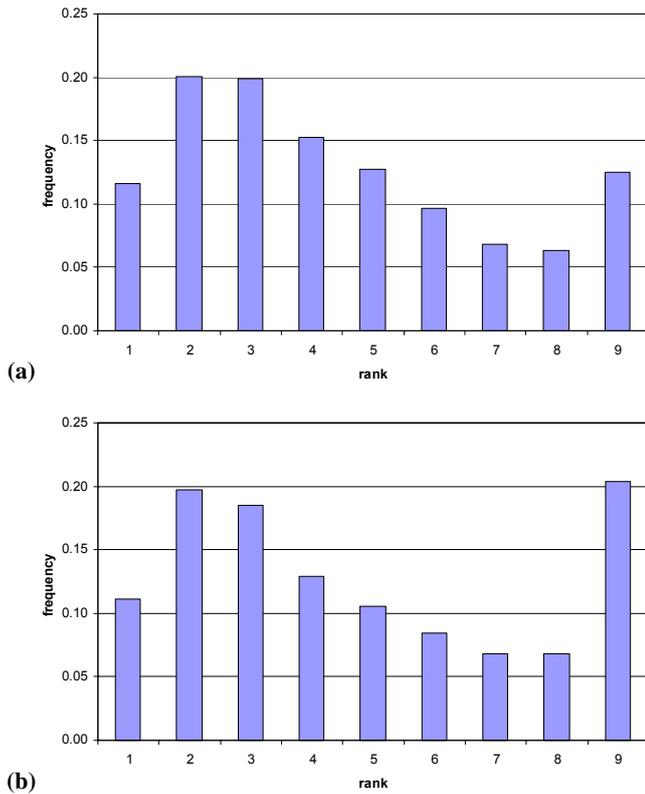
**Fig. 3.** Rank histogram of the observation with respect to the ensemble forecasts for average precipitation (a) and maximum precipitation (b) over the warning areas.

previously, is the Multimodel SuperEnsemble (Eq. 3) which produces a deterministic but calibrated value from a given ensemble.

### 3 Multimodel evaluation of precipitation

Precipitation fields are quite difficult to handle with post-processing techniques, although they are of great interest for operational NWP: the use of Multimodel technique on precipitation requires great care and we developed some simple correction to the basic method to obtain a better precipitation forecast.

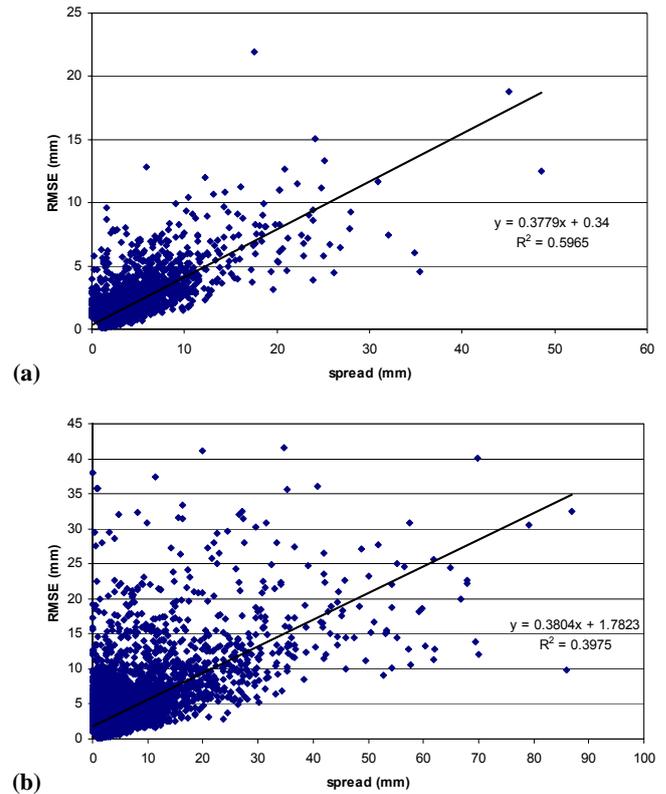
1. The weights can be calculated only once, on a sufficiently long training period, and then applied during



**Fig. 4.** Corrected rank histogram of the observation with respect to the ensemble forecasts for average precipitation (a) and maximum precipitation (b) over the warning areas.

the forecast time. We are using instead a moving window for the calculation of the weights: for each day, point and forecast time we evaluate the availability of models and then we calculated weights in the past with the given model number and dataset. This technique is more expensive than a fixed weight calculation, but gives better results and is more flexible for an operational use (Cane and Milelli, 2006).

2. In case of very scarce forecasted precipitation, it is possible to have a spurious bias from the method. Let us consider for example 8 models, 2 of them forecasting 1–2 mm for a given forecast time and the others 0 mm: looking at (Eq. 3), the result is a non-zero precipitation even if probably a zero precipitation should be the best forecast. We then developed a bias correction methodology based on contingency table evaluation during the training period. For each model we calculate the False Alarm Rate (FAR) and Hit Rate (HRR) and, treating them as independent probabilities, we evaluate the Multimodel FAR and HRR by multiplying the  $Z$  models having precipitation indices with the reverse indices of the other  $N-Z$  models:



**Fig. 5.** Ensemble mean Root Mean Square Error vs ensemble spread for average precipitation (a) and maximum precipitation (b) over the warning areas.

$$FAR_{MM} = \prod_{i=1}^Z FAR_i \cdot \prod_{i=1}^{N-Z} (1 - FAR_i) \quad (6)$$

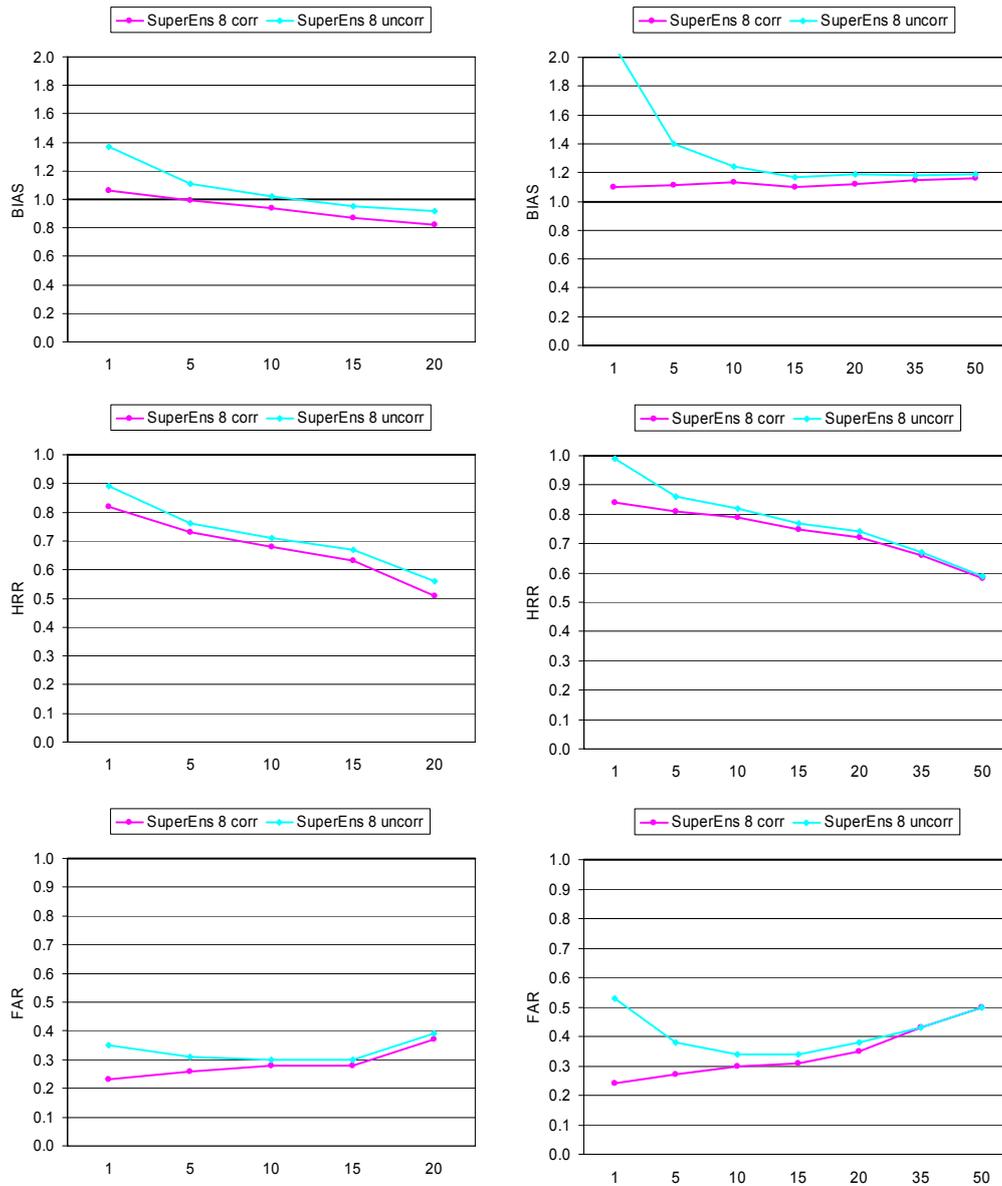
$$HRR_{MM} = \prod_{i=1}^Z HRR_i \cdot \prod_{i=1}^{N-Z} (1 - HRR_i) \quad (7)$$

The Multimodel SuperEnsemble and Ensemble are set to 0 if  $FAR_{MM} \geq HRR_{MM}$ , that is to say, if the probability of an incorrect rain forecast is higher than the probability of a correct forecast.

## 4 Results

### 4.1 Bias-corrected vs. uncorrected Multimodel

Firstly we compared the behaviour of the original SuperEnsemble with respect to the one with the correction introduced in Sect. 3. In Fig. 6 we report the verification indices (Wilks, 1995) for Bias-corrected and uncorrected Multimodel on average and maximum values, calculated in the period September 2006–August 2008 using 8 runs of the 4 models from +12 h to +36 h. The training period is obtained with a moving window of 365 days.

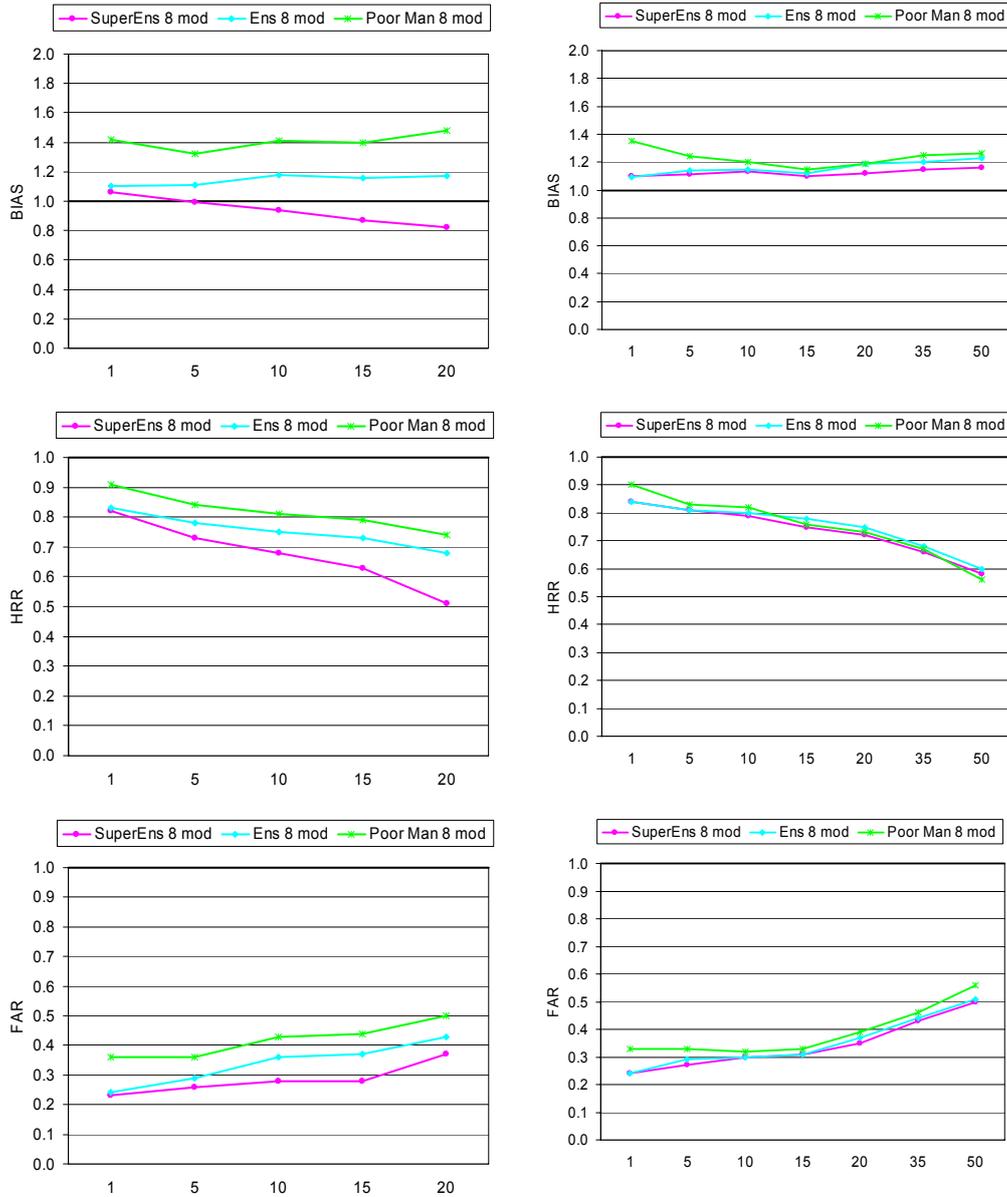


**Fig. 6.** Verification indices for Bias-corrected and uncorrected Multimodel as a function of the threshold (mm) for +12 h/+36 h forecasts. Average values (left) and maximum values (right) BIAS, HRR and FAR.

The bias correction has a neutral to positive impact: in particular, there is a significant improvement for low thresholds, with a reduction of the bias and of the FAR, while for high thresholds the two methods basically do not differ. This analysis is more evident for the maximum values. The impact on the lower thresholds is expected because the rain/no-rain limit has probably the largest uncertainty in the models (and this permits to have some improvement), while in case of a structured rain event, all the models can describe it (even though there are differences in the values) because the synoptic forcing is more clear.

#### 4.2 Multimodel SuperEnsemble, Ensemble, and Poor Man Ensemble

In the following step of the work we compared Multimodel SuperEnsemble (Eq. 3), Ensemble (Eq. 2), and Poor Man Ensemble (Eq. 1) in the period September 2006–August 2008 using 8 runs of the 4 models from +12 h to +36 h. The bias correction as described before (Sect. 3) is applied for each of them. As expected (Krishnamurti et al., 1999), in this case there is a clear difference between the three techniques (Fig. 7): Poor Man Ensemble shows a quite high BIAS, resulting in a strong FAR. SuperEnsemble and Ensemble show

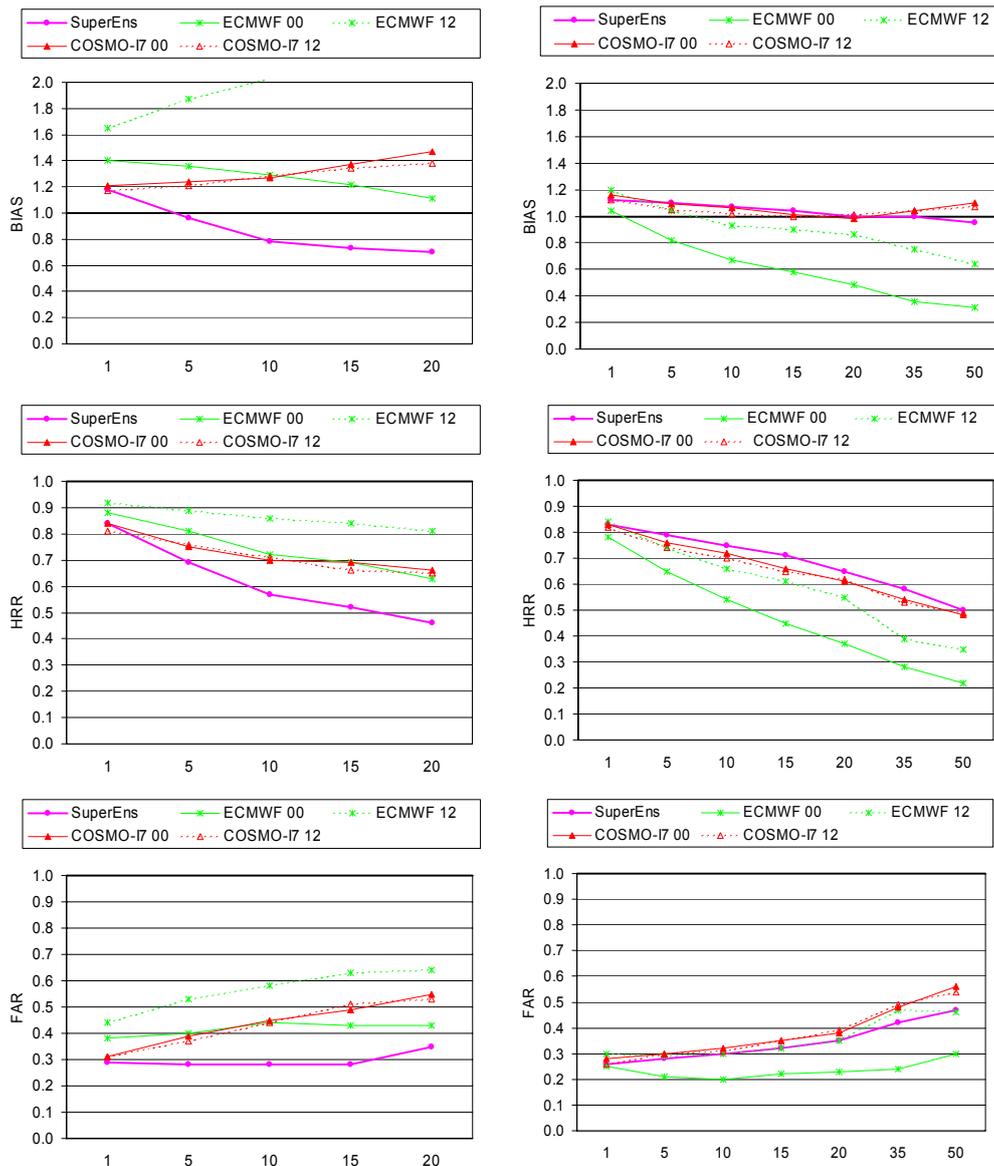


**Fig. 7.** Verification indices for Multimodel Ensemble, SuperEnsemble and Poor Man Ensemble as a function of the threshold (mm) for +12 h/+36 h forecasts. Average values (left) and maximum values (right) BIAS, HRR and FAR.

a very good reduction of the BIAS, very close to 1. There are few differences among SuperEnsemble and Ensemble for maximum values (we guess that the use of weights is not so helpful in this case because the bias reduction applied with the ensemble is the biggest contribution to the error reduction), while for average values the SuperEnsemble shows a good reduction of FAR, with a lower HRR with respect to Ensemble.

### 4.3 Operational forecasts: a comparison with the Direct Model Outputs (DMO)

Eventually, we analysed the results of our operational suite, obtained with the Multimodel SuperEnsemble applied on the ECMWF IFS global model and on the COSMO-17 limited area model in the period September 2006–August 2008 using 4 runs of the 2 models from +12 h to +36 h. Again, the contingency table correction has been applied. The post-processed data perform always better than the original models, with a sensitive bias reduction both on average and maximum values (Fig. 8). ECMWF shows a lower FAR for maxima, but this is clearly due to the strong under-biasing due



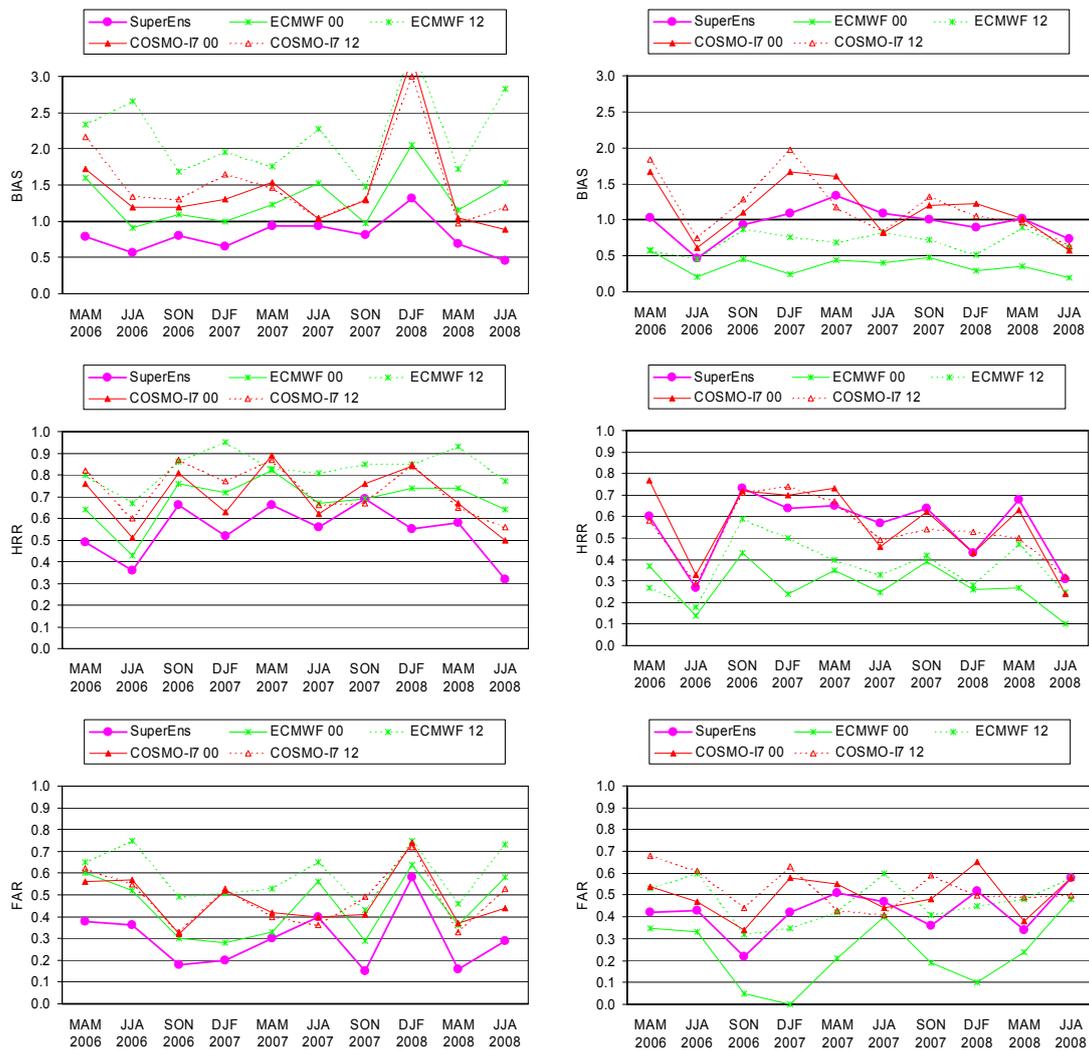
**Fig. 8.** Verification indices for Multimodel SuperEnsemble ECMWF IFS 00:00 UTC and 12:00 UTC, and COSMO-I7 00:00 UTC and 12:00 UTC runs as a function of the threshold (mm) for +12 h/+36 h forecasts. Average values (left) and maximum values (right) BIAS, HRR and FAR.

to the different horizontal resolution of the model in comparison to the others. ECMWF gives a smoother precipitation field and it is less prone to double-penalty errors that penalize the LAMs. For this reason the use of ECMWF together with the LAMs introduces a small added value also in maximum forecasts, because it usually underestimates values, but it is more reliable in time and space.

We could say that Multimodel SuperEnsemble is able to catch the best behaviour from each model: the lowest FAR for average values, with a BIAS slightly lower than 1; for maximum values a BIAS comparable with COSMO-I7, but better HRR and FAR.

Figure 9 shows the same comparison, for fixed significant thresholds, as a function of the season. SuperEnsemble is able to provide a stable-in-time performance, almost always better than the original models: in fact the BIAS is stably close to 1, the HRR is slightly lower for the average values and higher for maximum values and the FAR is significantly lower than the models (for maximum values only COSMO-I7 has to be taken into account due to the strong under-biasing of ECMWF).

The strong positive bias peak shown by the DMOs in Fig. 9 for DJF 2008 is probably due to a significant decrease of precipitation observed during the period with respect to the



**Fig. 9.** Verification indices for Multimodel SuperEnsemble ECMWF IFS 00:00 UTC and 12:00 UTC, and COSMO-I7 00:00 UTC and 12:00 UTC runs for +12h/+36h forecasts. Average values, threshold 10 mm (left) and maximum values, threshold 35 mm (right) BIAS, HRR and FAR.

climatology (around  $-40\%$  compared to the 1971–2000 average values from Arpa Piemonte archive); Multimodel SuperEnsemble is able to correct also this data effectively.

## 5 Conclusions and future perspectives

Given the fact that a probabilistic approach to precipitation forecast would be better than a deterministic one, because the quantification of the uncertainties would help the warning and decision-making process, we observed that the ensemble here considered is over-dispersed and uncalibrated. Therefore, the possible solutions are either the application of some calibration method (such as for instance BMA, see Raftery et al., 2005), or the construction of a super ensemble which is still deterministic. The latest approach is much easier than a BMA, hence it can be considered as a first step.

BMA itself it is not easy to implement in the case of precipitation forecast, because the estimation of the precipitation PDF is a very tricky task.

The Multimodel SuperEnsemble technique has been already applied in a number of different cases, frameworks and circumstances, but to the authors' knowledge, the applications to the precipitation field are quite scarce. The reason is obviously due to the peculiar characteristics of this variable. We have here introduced a correction method that corrects the QPF at least for low thresholds, reducing considerably the FAR.

We have then showed that this corrected SuperEnsemble has an added value with respect to the simple bias-corrected Ensemble, to the Poor-Man Ensemble and to the Direct Model Output (DMO) since, in general, the error is more stable with respect of threshold and time.

We are now working on Multimodel SuperEnsemble error evaluation methods, including bootstrap, and on ensemble calibration methods, in order to keep a probabilistic point of view. We are trying to evaluate the correct PDF of the Multimodel forecasts starting from the observed PDF conditioned to the forecasts, combining Multimodel SuperEnsemble and a BMA-like ensemble dressing: the first results are encouraging and we are planning to apply the so-found probabilistic precipitation forecasts as the input of our hydrologic chain, in order to evaluate the uncertainties in the discharge calculations. The use of finer spatial resolution (for example: at station locations) will be necessary for this application.

Moreover we are working on the downscaling of Regional Climatic Models with a Multimodel SuperEnsemble technique: in this case we use as training period the control runs and we apply the weights calculated in this period to the model scenarios.

*Acknowledgements.* This work is supported by the Italian Civil Protection Department. We wish to thank our colleagues Elena Oberto and Marco Turco for the useful discussions about the precipitation bias correction and Renata Pelosini for her support.

Edited by: S. Michaelides, K. Savvidou, and F. Tymvios

Reviewed by: two anonymous referees

## References

- Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Climate*, 9, 1518–1530, 1996.
- Cane, D. and Milelli, M.: Use of Multimodel SuperEnsemble Technique for Mountain-area weather forecast in the Olympic Area of Torino 2006, *Croat. Met. J.*, 40, 215–219, 2005.
- Cane, D. and Milelli, M.: Weather forecasts with Multimodel SuperEnsemble Technique in a complex orography region, *Meteorol. Z.*, 15(2), 1–8, 2006.
- Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, 129, 550–560, 2001.
- Krishnamurti, T. N., Kishtawal, C. M., Larow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved weather and seasonal climate forecasts from Multimodel Superensemble, *Science*, 285, 1548–1550, 1999.
- Krishnamurti, T. N., Kishtawal, C. M., Larow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Multi-model superensemble forecasts for weather and seasonal climate, *J. Climate*, 13, 4217–4227, 2000.
- Krishnamurti, T. N., Stefanova, L., Chakraborty, A., Vijaya Kumar, T. S. V., Cocke, S., Bachiochi, D. R., and Mackey, B.: Seasonal Forecasts of precipitation anomalies for North American and Asian Monsoons, *J. Meteorol. Soc. Jpn.*, 80(6), 1415–1426, 2002.
- Lovejoy, S., Duncan, M. R., and Schertzer, D.: Scalar multi-fractal radar observer's problem, *J. Geophys. Res.-Atmos.*, 101(D21), 26479–26491, 1996.
- Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using a Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, 133, 1155–1174, 2005.
- Wilks, D. S.: *Statistical Methods in the Atmospheric Sciences. An Introduction*, Academic Press, San Diego, 467 pp., 1995.
- Williford, C. E., Krishnamurti, T. N., Correa-Torres, R. J., Cocke, S., Christidis, Z., and Vijaya Kumar, T. S. V.: Real Time Multianalysis/Multimodel Superensemble Forecasts of Precipitation using TRMM and SSM/I Products, *Mon. Weather Rev.*, 131/8, 1878–1894, 2002.